CytoBase: An Electronic Medical Record for Cervical Cytology

Jack K. Golabek, P.Eng, Peter J. Brueckner, M.D., F.R.C.P.(C); AIM Inc.
Allan M. Seidenfeld, M.D., F.R.C.P.(C); Inscyte Corporation
Toronto, Ontario, Canada

A partnership of medical laboratories in Ontario, Canada has developed and made operational a centralized interactive database of cervical cytology (Pap) reports. This system automatically registers some 60,000 reports monthly, which are submitted electronically from geographically diverse sources. Patient histories are provided on-line to authorized pathologists to support ongoing diagnoses. A formalized coding system has been developed to represent elements of Pap reports in machine readable format, including diagnosis, clinical information about the patient, specimen adequacy, methods of specimen collection and follow-up recommendations. The system operates over a private network based on Internet standards. A review of operations, which commenced June 1, 1996, provides insights into applied EMR technology.

INTRODUCTION

In part, the impetus for developing an electronic medical record (EMR) for cervical cytology was provided by the Ontario Cervical Screening Collaborative Group, formed in 1993 to address the issue of a plateau in the incidence rate of cervical cancer in Ontario [1]. The collaborative group's recommendations included the formation of a screening program, the core of which would be a patient-oriented database of Pap, colposcopic and histologic findings.

In a proactive effort, six of Ontario's medical laboratories[1], which collectively process some 80% of the province's Pap tests, undertook to develop a centralized database to provide patient histories to pathologists to elevate the quality of diagnoses on current cases, ensure follow-up of positive results, and to support screening programs. The software components of this database are called CytoBase.

At the core of CytoBase is an Oracle 7™ RDBMS running under UNIX on a DEC Alpha 2000 server. All operational server software was written in 'C' while client applications were written in Visual Basic™ and PowerBuilder™ for Windows 3.11 and Windows® 95 platforms. The network protocol is TCP/IP and connections are currently accomplished using point-to-point protocol (PPP) over public carriers. An Internet web gateway is also being contemplated, pending resolution of certain policy issues. CytoBase is designed to interconnect with a variety of stakeholders as shown below.

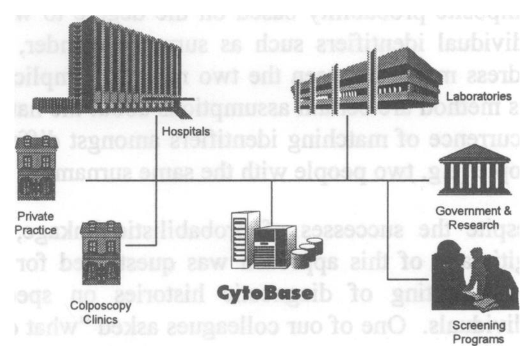

Figure 1: Stakeholder Community

Reports are submitted by automated means with a minimum of human intervention. Transmissions use HL7 (v2.2) messaging standard. Prior to submission, source laboratories also translate the content of their reports to a common coding system, thereby removing any ambiguities that may otherwise result from a multiplicity of reporting styles.

The system began operations on June 1, 1996 and is now being expanded to capture relevant histology and colposcopy reports as well. This project continues to address many scientific, technical and policy issues related to applied EMR technology, including:

- patient identification
- nomenclature, classification, and coding of data
- quality assurance
- security and confidentiality
- fulfillment of stated goals

[1] Dynacare Laboratories, Excel Bestview Medical Laboratories, FML Laboratories, Gamma North Peel Laboratory, MDS Inc., Medchem Laboratories

1091—8280/97/$5.00 © 1997 AMIA, Inc.

A discussion of these issues in a production environment forms the focus of this paper.

PATIENT IDENTIFICATION

Unique patient identification is an issue with non-trivial consequences in applied EMR technology. In a database architect's dream, each person would be identified by a single, unique and perfectly reliable identifier. But given our human desire for freedom (and hence a certain level of anonymity) it is doubtful if such a system will ever be realized. Consequently, EMR technology must rely on analytical methods to resolve patient identity.

Probabilistic linkage is one method that has been used in many epidemiological applications [2]. This method determines the likelihood that two patient records represent the same person by calculating a composite probability based on the degree to which individual identifiers such as surname, gender, and address match between the two records. Implicit in this method are certain assumptions about the natural occurrence of matching identifiers amongst different people (e.g. two people with the same surname).

Despite the successes of probabilistic linkage, the legitimacy of this approach was questioned for on-line reporting of diagnostic histories on specific individuals. One of our colleagues asked "what does it really mean to say that an individual is 97% like another"? In other words, can decisions about a patient's treatment be defended if they are based on a probabilistic interpretation of the patient's history? Either the patient had the Pap test or she didn't.

We decided to use a deterministic network to cross-reference patient records based on three identifiers: health insurance number, surname, and date of birth, all of which are provided in over 90% of Pap reports. When a report is submitted, the registration process first searches for an exactly matching patient in the patient registry. If one is found, the inbound report is linked to that patient record, otherwise the system creates a new patient record. A process is then invoked to cross-reference the new record with existing patient records, and labeling each link according to the following classes:

A: Exact match (used for manual upgrades only)
B: Mismatch on surname
C: Mismatch on date of birth
D: Mismatch on health insurance number
E: Mismatch on surname and date of birth
... and so on.

Figure 2 illustrates how patient records are cross-referenced using this method. In this example, six reports have been received by the system (over a period of time), resulting in four patient records and the cross-references shown. Note that each patient record is linked to the remaining three.

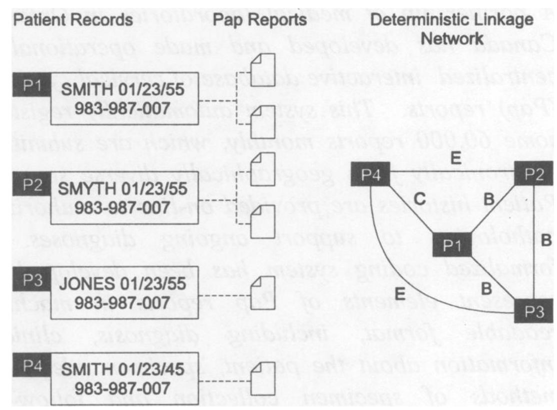

Figure 2: CytoBase Patient Linkage

To retrieve a patient history requires a surname, date of birth and health insurance number to be submitted. In addition, the user specifies a minimum link class to limit how far into the network the system will search for reports. When query results are returned, the link class and patient identifiers display with each report indicating unambiguously any discrepancies in patient matching. Determining which reports enter into treatment decisions is then up to the practitioner. Note that regardless of the patient originally searched for, all appropriate reports will be retrieved.

NOMENCLATURE, CLASSIFICATION AND CODING OF DATA

The benefits of using standardized codes and nomenclatures to represent medical data in an EMR are well established [3]. An investigation into reporting practices showed that in our environment all Pap reports could be described by the following superset of elements:

- clinical information about the patient as at the time of the test (e.g. date of last menstrual period, cycle duration, therapy, etc.)
- a detailed description of specimen adequacy
- the formal diagnosis
- pathologist's comments and recommendations for follow-up testing or treatment
- method of specimen collection

- method of slide preparation
- test method (traditional or automated)

A machine readable representation was required for each of these elements to ensure a broad scope of application for the CytoBase system.

We reviewed both SNOMED III and ICD-9 as candidates for classifying diagnoses and concluded that each was capable of expressing certain diagnostic details. However, the stakeholders in our community favored the Bethesda System of Pap reporting [4]. This system classifies Pap results into categories ranging from "no abnormal cells" to "invasive carcinoma", but does not define any standardized codes for these classes.

We constructed a generalized coding method to permit machine manipulation of all elements of the Pap report. This coding system, which we refer to as the *CytoBase Abstract Coding Syntax*, consists of numbers separated by periods, each number representing an element of information arranged in a hierarchical tree structure. To illustrate, the code "4.6.2.2" represents the following elements of information:

4: [diagnosis]
    6: [atypical glandular cells]
        2: [endometrial origin]
            2: [favor pre-neoplastic]

Internally, code "4.6.2.2" is cross-referenced to a dictionary of terms and is reported literally as *"Atypical glandular cells of endometrial origin are seen, query pre-neoplastic."* This separation of codes from reporting phraseology also enables CytoBase to interact with users in multiple languages.

Certain codes also allow free text annotation to further qualify the data. For example, code "1.8", which represents [clinical history] [weeks post-partum] requires a numeric modifier to quantify the number of weeks post-partum.

This coding method is not intended to be a proposal for standardized Pap reporting. Rather it is a necessary abstraction that casts all elements of a Pap report into a machine readable format. This enables the EMR to automatically tabulate statistics and to initiate specific actions based on report content. For example, when a Pap report is registered, the system automatically tags the patient record with a follow-up date calculated on the basis of diagnostic class and pathologist's recommendations.

Integrating SNOMED, the ICD or the UMLS with the CytoBase Abstract Coding Syntax is a compelling avenue for future work. Additionally, we are investigating how our coding syntax may contribute to the LOINC initiative [4] for standardizing the identification of laboratory observations.

QUALITY ASSURANCE

A key concern in the operation of an EMR is how to manage the flow of data to ensure accuracy, reliability, and timely delivery of information to the appropriate audience. If an EMR is intended to support providers' decision making activities, then these issues are of paramount importance. In a high volume system such as CytoBase, where human intervention is minimized in favor of data throughput, automated quality control mechanisms are essential.

Having established a machine readable format for submitted data, automated screening for certain elements of quality can be readily implemented. For example, examination of the codes submitted within a report can screen out inconsistencies by identifying conflicting code combinations. Similarly, screening of patient identifiers can identify reports containing insufficient information, or instances where non-exact patient linking has occurred.

In the CytoBase system, an automated feedback mechanism routinely notifies laboratories about inconsistencies in submitted reports, missing data, and non-exact patient linkage. The registration process analyzes inbound reports and automatically prepares acknowledgment messages detailing any problems encountered, and the disposition of the report (i.e. registered or rejected). In most cases, the source laboratory can rectify these problems and re-submit the affected reports.

Although more sophisticated methods can be implemented (e.g. linking with external government databases to resolve the identity of registered patients), the key factor to success remains the willing participation of source laboratories in the quality control effort.

In implementing the CytoBase system, we have witnessed an increased awareness of quality issues on a grander scale than previously realized. Indeed, the participating laboratories have formed an independent "user group" to regularly review their involvement and interaction with the system. Suggestions have been made about adopting a common nomenclature for Pap reporting and CytoBase is viewed as an added layer of quality assurance. Consequently, it appears that EMR's can contribute positively to forming collaborative quality assurance initiatives amongst competing market entities.

SECURITY AND CONFIDENTIALITY

While electronic medical records offer significant benefits for improved health care, it has been stressed that these should not come at the expense of a patient's right to privacy and confidentiality [5]. Furthermore, it has long been identified that in the medical domain, the problem with security and confidentiality lies not in technology, but in the lack of cohesive policies [6]. Consequently, in designing the CytoBase system our attention was first focused on developing policies that would foster awareness and compliance with confidentiality issues amongst all stakeholders. Then, technical security provisions were developed to support and actualize the substance of these policies.

Guiding Principles

Our stakeholder community adopted the following principles as the foundation on which to build a security and confidentiality program:

1. An individual has the right to privacy of personal health information.

2. Data about a patient's state of health (problems, diagnoses, test results, etc.) are the property of the patient. The collection and documentation of this information is entrusted to practitioners.

3. To provide health care services of the highest quality and efficacy, practitioners require access to patient identified medical information on a need to know basis.

4. Information provided to practitioners must be highly reliable, accurate and readily accessible if it is to be useful.

With the above principles in mind, specific procedures and protocols were developed to ensure that the functional and operational goals of CytoBase could be achieved in concert with confidentiality concerns. These included the development of formal licensing agreements between stakeholders and Inscyte Corp., the governing body of CytoBase.

Interaction with CytoBase is currently restricted to authorized laboratories and the Ontario Cancer Treatment and Research Foundation (OCTRF). Each laboratory is required to execute a formal *Reporting Site License Agreement* which outlines in detail the site's obligations and responsibilities with respect to access and utilization of information. In addition, at each reporting site, every individual who will have access to CytoBase data must sign a formal *Confidentiality and Non-Disclosure Agreement* explaining the nature of CytoBase information, appropriate use, and remedies available to affected parties in case of breach. To promote continuing awareness, these agreements must be renewed annually.

The primary technical security mechanism is the electronic authentication of a security ticket, comprised of five elements that must individually and collectively pass the authentication process before access rights are granted. These elements are: a site identifier, license number, user name, password and database identifier.

The second major security provision is the audit system, which comprises four logs that record all events and actions initiated against the database at the network, operating system, file system and database levels. These logs are used to record attempted violations and support disciplinary actions, but are also essential to quality control monitoring.

FULFILLMENT OF STATED GOALS

CytoBase was developed to fulfill two major health care goals: to provide information facilitating a 50% reduction in the mortality rate from cervical cancer in Ontario by the year 2005, and to track patterns in the utilization and efficacy of cervical cytology testing on a province-wide scale for policy development.

With respect to the former, CytoBase is proving to be an excellent tool for tracking individual patients' health status regardless of where the patient is tested or where health care services are obtained. CytoBase can automatically identify patients who are in need of treatment or repeat testing. This capability can contribute greatly to a proposed province-wide screening program by ensuring that abnormalities are identified in the early stages, and that follow-up activities are not overlooked. Of course, this applies only to women registered in the system, and other methods will be required to reach women who have not had Pap tests done.

Table I shows the distribution of women registered in the CytoBase system to January 21, 1997. Already, registered patients account for 4.23% of the female population, even though not all laboratories had been connected during this time. As more laboratories and hospitals join the project, we expect this figure to increase dramatically. Our goal is to involve all testing sites, and thereby capture 100% of tested patients.

Table I: Registered Patients as at January 21, 1997 (where date of birth was reported)

Age Group	Number Registered	Female Population*	% of Cohort Registered
0-9	37	712,752	0.01%
10-19	11,042	656,028	1.68%
20-29	57,231	830,551	6.89%
30-39	63,377	936,044	6.77%
40-49	43,720	752,875	5.81%
50-59	26,584	489,949	5.43%
60-69	15,871	470,101	3.38%
70+	7,927	494,015	1.60%
Total	225,789	5,342,315	4.23%

* Division of Preventive Oncology, OCTRF: post census estimates for 1993 Ontario population, Toronto 1996.

From June 1, 1996 to January 21, 1997, CytoBase registered 229,466 reports on 225,914 women, 4,807 requesters and 79 pathologists. The system continues to register some 60,000 reports per month. In the first 8 months of operation, only 3% of reports in the system represented a repeat test on a patient. Since many women have Pap tests done annually, we expect this figure to increase quickly. It is also encouraging to note that of all the secondary reports, 94% matched exactly to an existing patient record, 4% were linked under class "B", and 2% class "C".

Statistics compiled to January 21, 1997 already provide useful information on the distribution of diagnoses by patient age, the overall quality of specimen collection, and recommendations being made vis-à-vis diagnostic severity. As historical data accumulate it is evident that CytoBase will provide highly valuable information on the incidence of abnormalities and the progression of cervical cancer. We expect these data to support policy formulation regarding appropriate testing intervals (utilization) and follow-up actions.

Based on our experience, we believe that effective EMR technology can be successfully implemented provided certain key factors are realized. Foremost is the dedication of stakeholders to achieving a clear set of goals. Adoption of a common data format and nomenclature, and casting these in machine readable format is essential. Formulating operating policies at the outset allows smoother integration into daily routines. And finally, adherence to engineering and computer standards ensures the technology works.

More information about this project and a complete listing of database statistics may be found on the Internet at *www.inscyte.org* E-mail may be sent to *aim@servtech.com*. Mail may be addressed to the authors at:

AIM Inc., 10 Gateway Boulevard, Suite 340
Toronto, Ontario, CANADA
M3C 3A1

Acknowledgments

We wish to acknowledge Dr. Aileen Clarke, the Ontario Cancer Treatment and Research Foundation, and the Ontario Cervical Screening Collaborative Group for their help and support of this project. We also thank all member laboratories for their contributions to the success of this project.

References

1. Holowaty E.J., Marrette L.D., Fehringer G. Cancer Incidence in Ontario: Trends and Regional Variations. OCTRF, 1995: 36-37.
2. Matthew A. Jaro. Probabilistic Linkage of Large Public Health Data Files. Statistics in Medicine, 1995: 14; 491-498.
3. Board of Directors of the AMIA. Standards for Medical Identifiers, Codes, and Messages Needed to Create an Efficient Computer-stored Medical Record. JAMIA, 1994: 1; 1-7.
4. Regenstrief Institute, Indianapolis, IN 46202. Laboratory Observation Identifier Names and Codes. Loinc@regenstrief.iupui.edu
5. Samuel Broder. The Bethesda System for Reporting Cervical/Vaginal Cytologic Diagnosis - Report of the 1991 Bethesda Workshop. JAMA 1992: 267; 1892.
6. Randolph C. Barrows Jr., Paul D. Clayton. Privacy, Confidentiality and Electronic Medical Records. JAMIA, 1996: 3; 139-148.